MATH 343 LEC 1

Recall $\underbrace{f(\theta_1,\dots\theta_K|\vec{X})}_{\text{posterior}} \propto \underbrace{f(\vec{X}|\theta_1,\dots\theta_K)}_{\text{likelihood}} \underbrace{(\theta_1,\dots\theta_K)}_{\text{prior}}$

We are ultimately interested in $f(\theta_i|\vec{X})$ possibly for all $i=1,\dots,K$

if not $\propto$ a PDF/PMF which is known and easily sampled from, exact inference is not possible, so we use computational techniques to approximate inference.

Sampling $\{\theta_{i,1},\theta_{i,2},\dots\}$ iid from $f(\theta_i|\vec{X})$ gives us the ability to estimate $\hat{\theta}_i$'s, CR's, hypothesis tests

One such method is called the systematic sweep "Gibbs Sampler"

Step 0: Initialize $\vec{\theta}_0=[\theta_{0,1},\theta_{0,2},\dots,\theta_{0,K}]$

Step 1: Draw $\theta_{1,1}$ from $f(\theta_1|\vec{X},\theta_{0,2},\dots\theta_{0,K})$

Step 2: Draw $\theta_{2,1}$ from $f(\theta_2|\vec{X},\theta_{1,1},\theta_{0,3},\dots,\theta_{0,K})$

⋮

Step K: Draw $\theta_{K,1}$ from $f(\theta_K|\vec{X},\theta_{1,1},\theta_{1,2},\dots\theta_{1,K-1}) \Rightarrow \vec{\theta}_1=[\theta_{1,1},\dots,\theta_{1,K}]$

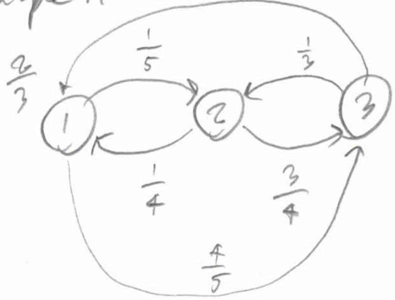
Step K+1: Repeat steps 1-K many times until convergence

Proof (not covered!) for derivation... whole class on Markov chains! 623

Def: A 1st order Markov Chain on space $\mathcal{X}$ is a seq of rv's $X_0, X_1, X_2, \dots$ s.t. for $A \subseteq \mathcal{X}$

$P(X_{t+1}\in A | X_0,\dots X_t) = P(X_{t+1}\in A|X_t)$ i.e. only the previous realization matters! Only "sees" one step in the past. memory loss!

Example in discrete case, one variable X, $S_X = \{1,2,3\}$. Chain $X_0, X_1, X_2, \ldots$



If at 1, you can transition to 2 or 3.

If at 2, ... 1 or 3,

. . . . 3, ... 1 or 2

let A be the transition matrix

$$\begin{array}{cc} & \text{start} \\ & \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \text{step} \begin{array}{c} 1 \\ 2 \\ 3 \end{array} & \begin{bmatrix} 0 & \frac{1}{4} & \frac{2}{3} \\ \frac{1}{5} & 0 & \frac{1}{3} \\ \frac{4}{5} & \frac{3}{4} & 0 \end{bmatrix} \end{array} = A$$

Usually defined as A^T ... but who cares...

let $\vec{p}_0$ be the starting position e.g. $\vec{p}_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ means $X_0 = 1$

$\vec{p}_1$ is prob distr of first step i.e. $\vec{p}_1 = A\vec{p}_0 = \begin{bmatrix} 0 \\ \frac{1}{5} \\ \frac{4}{5} \end{bmatrix}$

$\vec{p}_2$ . . . . second step . . . $\vec{p}_2 = AA\vec{p}_0 = A\vec{p}_1 =$

$\vdots$

$\vec{p}$ is the steady distr. $\vec{p} = \lim_{n\to\infty} A^n \vec{p}_0 = A\vec{p}$, i.e. the eigenvector with eigenvalue $=1$ scaled to sum to one.

Invariant distr is:

$$\underset{X}{f}(x) = \int_{\mathcal{X}} \overbrace{f_{X_{t+1}|X_t}}^{\text{transition kernel}}(x,y)\, f_X(y)\, dy \quad \forall t$$

i.e. average over all states of previous step

Ignoring technical details... for Markov Chain $X_0, X_1, X_2, \ldots$ with invariant distr. $f_X(x)$,

Thm: $\forall x_0 \in \mathcal{X}$, $f_X(x) = \lim_{t\to\infty} \int_{\mathcal{X}} \prod_{i=0}^{t} f_{X_{i+1}|X_i}(x,y)\, f_0(x_0)\, dy$

$\forall f_0$ which is a valid starting distr.

This is in one dimension, the Gibbs Sampler is K dimensions. It is a Markov Chain!

If we prove steps 1–K is a transition kernel and $f(\theta_1, \ldots \theta_K \mid \vec{x})$ is the invariant distr, then it will converge to the invariant distr. regardless of starting distr $\vec{\theta}_0$

$$f(\theta_1 = \theta_{t+1,1}, \ldots \theta_K = \theta_{t+1,K} \mid \vec{x}) \overset{?}{=} \int_{\theta_{t,1}} \int_{\theta_{t,2}} \cdots \int_{\theta_{t,K}} (\text{steps } 1, \ldots, K)\, \overbrace{f(\theta_1 = \theta_{t,1}, \ldots \theta_K = \theta_{t,K} \mid \vec{x})}^{\text{starting position}}\, d\theta_{t,1} \ldots d\theta_{t,K}$$

$$f(\theta_{t+1,1}, \ldots \theta_{t+1,K} \mid \theta_{t,1}, \ldots, \theta_{t,K}, \vec{x})$$

$= f(\theta_{t+1,1} \mid \theta_{t,2}, \ldots \theta_{t,K}, \vec{x}) \cdot$ Step 1

$f(\theta_{t+1,2} \mid \theta_{t+1,1}, \theta_{t,3}, \ldots \theta_{t,K}, \vec{x}) \cdot$ Step 2

$\vdots$ $\vdots$

$f(\theta_{t+1,K} \mid \theta_{t+1,1}, \ldots \theta_{t+1,K-1})$ Step K

$f(A,B) = f(A|B)\, f(B)$

$$= \int_{\theta_{t,2}} \cdots \int_{\theta_{t,K}} (\text{steps } 2,3,\ldots,K) \int_{\theta_{t,1}} f(\overset{B}{\theta_{t+1,1}} \,|\, \overset{A}{\theta_{t,2},\ldots,\theta_{t,K}}, \vec{x})\, f(\overset{A}{\theta_{t,2},\ldots,\theta_{t,K}})\, f(\theta_{t,1} \,|\, \theta_{t,2},\ldots,\theta_{t,K}, \vec{x})\, d\theta_{t,1}\, d\theta_{t,2} \ldots d\theta_{t,K}$$

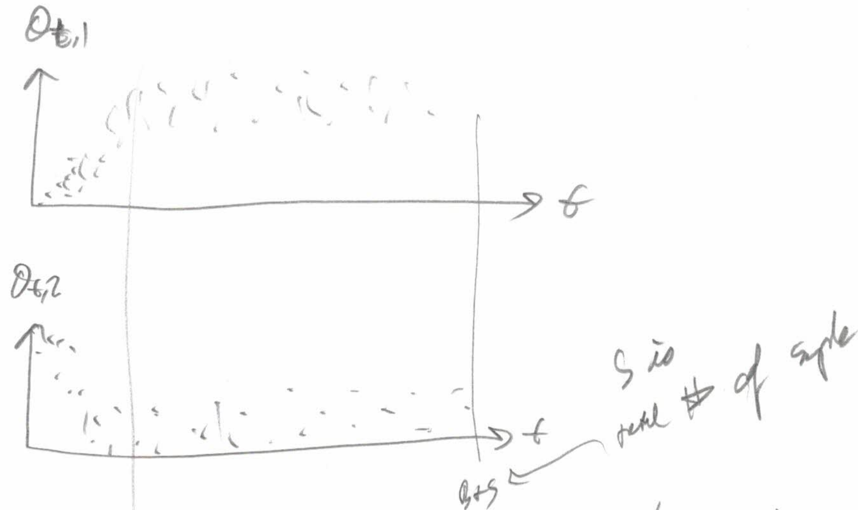
$$= \;,\;,\;,\;,\;,\;,\; f(\theta_{t+1,1}, \theta_{t,2},\ldots,\theta_{t,K}) \underbrace{\int_{\theta_{t,1}} f(\theta_{t,1} \,|\, \theta_{t,2},\ldots,\theta_{t,K})\, d\theta_{t,1}}_{1}\, d\theta_{t,2} \ldots d\theta_{t,K}$$

repeat K times...

$$= f(\theta_{t+1,1}, \ldots \theta_{t+1,K} \,|\, \vec{x}) \quad \blacksquare$$

Necessary: all $f(\theta_{t,j} \,|\, \vec{\theta}_{t,-j}, \vec{x}) > 0$

Practical Concerns... ① when does it converge? $t \to \infty$...

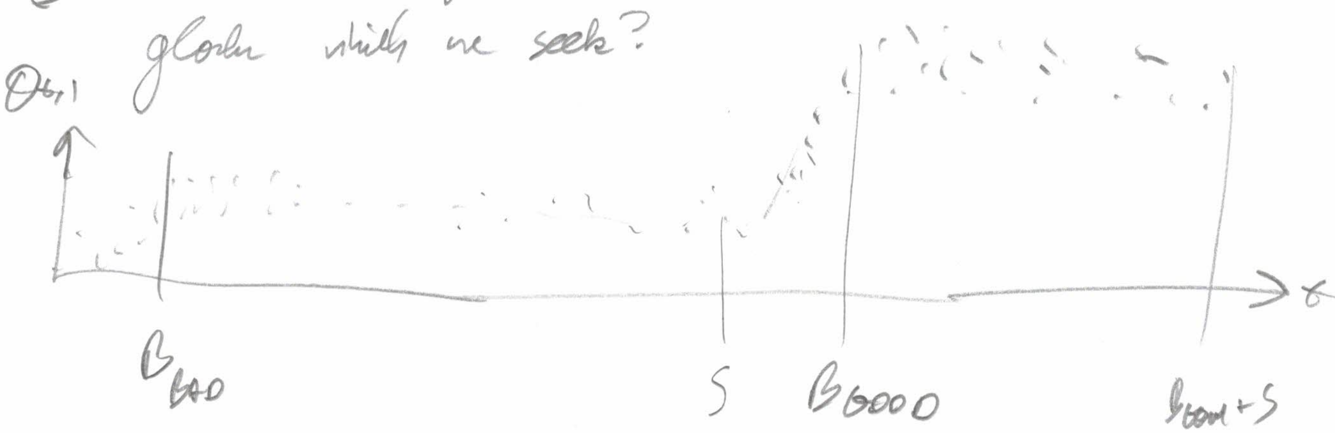


B means "burn-in" which is done with visual inspection

there are advanced methods to "test". Keep only samples after B

$$\langle \vec{\theta}_1, \vec{\theta}_2, \ldots \rangle \xrightarrow{\text{BURN}} \langle \vec{\theta}_B, \vec{\theta}_{B+1}, \ldots \rangle$$

② Did we converge to a local point of convergence (not the global which we seek?



Partial Solution: start the sampler from multiple locations.

Only a problem if K large

(3) $\vec{\theta}_{t+1}$ depends on $\vec{\theta}_t$. But we want to sample iid

Recall $\rho = Corr[X,Y] := \dfrac{Cov(X,Y)}{SE(X)\,SE(Y)} = \dfrac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{Var(X)\,Var(Y)}}$

Method of Moments estimator $r = \dfrac{s_{X,Y}}{s_X s_Y} = \dfrac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum (x_i-\bar{x})^2}\sqrt{\sum (y_i-\bar{y})^2}}$

Here we are correlating $\vec{\theta}_{t+1}$ to $\vec{\theta}_t$ and $\vec{\theta}_{t-1}, \vec{\theta}_{t-2}, \dots$ etc

let

$$r_{\ell_1} := \frac{\sum_{t=B}^{B+S-1} (\theta_{t,j} - \bar{\theta}_j)(\theta_{t+1,j} - \bar{\theta}_j)}{D_j} \qquad D_j := \sum_{t=B}^{B+S} (\theta_{t,j} - \bar{\theta}_j)^2$$

denominator approx correct

$$r_{\ell_2} := \frac{\sum_{t=B}^{B+S-2} (\theta_{t,j} - \bar{\theta}_j)(\theta_{t+2,j} - \bar{\theta}_j)}{D_j}$$

$\vdots$

$$r_{\ell} := \frac{\sum_{t=B}^{B+S-\ell} (\theta_{t,j} - \bar{\theta}_j)(\theta_{t+\ell,j} - \bar{\theta}_j)}{D_j}$$

Then choose lag ℓ_0 ..

Keep every ℓ_0 of the samples